

**Plans complexes, variances robustes et poids de rééchantillonnage.  
Problèmes et solutions vus par un chercheur**

**Benoît Laplante**

Centre interuniversitaire d'études  
démographiques

INRS Urbanisation, Culture et Société

# Plna de la conférence

- Statistique, statistique d'enquête et statistique de modélisation
- Plans complexes et le problème de la variance des estimés
- Deux programmes « versatiles » qui permettent d'utiliser les poids de rééchantillonnage avec Stata
- Exemples avec Stata

# Les mondes de la statistique que doit fréquenter tout chercheur

- La statistique « tout court »
- La statistique d'enquête
- La statistique de modélisation

# La statistique « tout court »

- Décrire les caractéristiques de l'État
- Descriptif
- Non probabiliste
- Au sens premier, la statistique sociale est descriptive, populationnelle et non probabiliste.

# Statistique d'enquête

- La population est finie.
- On cherche à mesurer une caractéristique dont la valeur précise existe nécessairement.
- En principe toute l'imprécision vient de l'erreur d'échantillonnage.
- On peut réduire cette imprécision en tirant des échantillons tirés au sein de sous-population relativement homogènes.
- On se trouve ainsi à décomposer l'erreur d'échantillonnage.

# Statistique d'enquête

- Le but est de mesurer et de décrire
- Au mieux, on décrira des sous-populations définis par les catégories d'une ou plusieurs variables
- La théorie des probabilités sert à modéliser l'imprécisions de la mesure due à l'échantillonnage, jamais les processus sociaux.

# Statistique de modélisation

- On présume que le monde a été créé par un modèle dont une composante est déterministe et l'autre et aléatoire.
- Toute la dispersion est générée par la composante aléatoire du modèle.
- On cherche à estimer les paramètres de ce modèle.
- On présume que l'échantillon dont on dispose est tiré de manière aléatoire simple de la population infinie que peut générer le modèle.

# Modèle de la super-population

- La population finie est générée par le modèle.
  - La population finie est un échantillon tiré au sein de la population infinie que peut générer le modèle.
- L'échantillon est tiré de cette population finie.
  - L'échantillon est donc lui-même tiré dans un échantillon.
- Les estimés ponctuels calculés à partir de cet échantillon (qui sont des estimés des paramètres de la population finie) sont également des estimés des paramètres du modèle.
- Les estimés des variances de ces estimés doivent être calculés en tenant compte du plan d'échantillonnage au sein de la population finie.

# Plans complexes

- Pourquoi:
  - Absence de registre de la population dont on pourrait tirer des échantillons
  - Coût
  - Réduire l'imprécision des estimés

# Plans complexes

- **Strates**
  - Réduisent l'imprécision des estimés
  - Décomposition de la variance analogue à celle de l'analyse de variance
- **Grappes**
  - Réduisent les coûts
  - Réduisent la puissance de l'échantillon parce que celle-ci dépend du nombre d'unités d'échantillonnage indépendantes.

# Plans complexes

- On cherche à obtenir des strates dont chacune est homogène du point de vue social et économique.
- Le plan de l'Enquête sur la population active contient ainsi près de 300 strates.
- À Statistique Canada, souvent une seule grappe par strate.

# Estimation de la variance de estimés

- Échantillon aléatoire simple
  - Théorème de la limite centrale
  - Solution algébrique analogue pour les modèles linéaires

$$\hat{\sigma}_{\bar{X}}^2 = \frac{\hat{\sigma}_X^2}{n-1}$$

- Échantillon à plan complexe
  - Plus compliqué...

## Estimation de la variance de estimés

- Correction de Kish.
- Méthode dite de la linéarisation, de Taylor, du sandwich, de Huber ou de White.
- Méthodes de rééchantillonnage.

# Correction de Kish

- On multiplie la matrice des variances et des covariances obtenue en présumant que l'échantillon est aléatoire simple par un estimé de l'effet de plan moyen.
- Correction approximative utile lorsque
  - On met au point un modèle et on veut tenir compte de l'effet de plan sans utiliser un méthode de rééchantillonnage.
  - On utilise une enquête de StatCan et on dispose d'un estimé de l'effet de plan mais pas de poids de bootstrap...

# Correction de Kish

Effets de plan	Secteur géographique	Effet du plan
Enquête sociale générale 2001  Source: Guide de l'utilisateur, p. 25	Canada	1,38
	Terre-Neuve	1,34
	Île-du-Prince-Édouard	1,22
	Nouvelle-Écosse	1,27
	Nouveau-Brunswick	1,84
	Québec	1,23
	Ontario	1,23
	Manitoba	1,21
	Saskatchewan	1,19
	Alberta	1,28
Colombie-Britannique	1,26	
Région Atlantique	1,56	
Région des Prairies	1,37	

# Méthode dite de la linéarisation

Estimateur de la variance de la moyenne:

$$\hat{\sigma}_{\bar{X}}^2 = \sum_{h=1}^L (1 - f_h) \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left( \bar{X}_i - \bar{\bar{X}}_h \right)^2$$

où  $L$  est le nombre de strates dans la population,  $n_h$  est le nombre d'unités primaires d'échantillonnage dans chaque strate  $h$ , et  $f_h$  est le rapport du nombre de grappes de cette strate et du nombre total de grappes au sein de cette strate.

## Pourquoi la méthode du bootstrap?

- Parce que la méthode de Taylor ne peut pas être utilisée lorsqu'on ne dispose que d'une seule grappe au sein d'une strate.
- Parce que StatCan n'inclut jamais la grappe et la strate auxquelles appartiennent les individus afin de ne pas permettre leur identification.
- Parce que la méthode dite du « jackknife » demande que l'on tire autant d'échantillons qu'on a d'unités d'échantillonnage.

# Méthode du bootstrap

- Tirer plusieurs échantillons **de grappes** au sein de l'échantillon original.
- Recalculer les poids finaux dans chacun de ces échantillons de manière à ce que chacun soit un échantillon isomorphe et probabiliste de la population finie.
- Estimer le modèle à partir de chacun de ces échantillons.
- Calculer la variance et la covariance de ces estimés.

# Méthode du bootstrap

- Le tirage des échantillons et le calcul des poids sont des opérations compliquées.
- Elles sont faites une fois pour toutes par le personnel de StatCan qui crée un ensemble d'échantillons rééchantillonnés.
- Les poids de rééchantillonnage sont l'outil qui permet à l'utilisateur d'utiliser ces échantillons:
  - chaque jeu de poids recrée un des échantillons rééchantillonnés par Statcan.

# Usage de la méthode du bootstrap

- Avec SPSS et SAS
  - BOOTVAR
- Avec SAS
  - MacBoot8

# Usage de la méthode du bootstrap

- Avec Stata

- -btstrap- de Darren Lauzon
- -bswreg- d'Emmanuelle Piérard, Neil Buckley et James Chowhan
- -bs4rw-, de Jeff Pitblado, un employé de Stata Corp.

# Pourquoi -bts- et -stbts-

- Parce que les autres programmes écrits pour Stata soit
  - Sont difficiles à comprendre.
  - Ne permettent pas d'ajouter des instructions supplémentaires lorsque nécessaire.
    - V.g. orthogonalisation
  - Ne permettent pas d'utiliser les instructions de la série -st-.

# Le cœur de -bts-

```
matrix b0 = e(b)'
matrix V = (b0 - b0)*(b0 - b0)'

foreach wname in `rw' {
  qui `cmd' `varlist' `if' `in' , `cmdops'
  matrix V = V + (e(b)' - b0)*(e(b)' - b0)'
  local B = `B' + 1
  if mod(`B',10)==0 di in gr "On a utilisé le `B'ème jeu de poids."
}

matrix b0 = b0'
matrix V = (`r'^`B')*V

ereturn post b0 V, dof(`dof')
ereturn display, level(`level') eform(`eform')
```

## -bts-

```
syntax varlist(numeric) [if] [in], cmd(string) [cmdops(string)] ///  
    PWeight(varname numeric) ///  
    rw(varlist numeric) [r(integer 1)] ///  
    [dof(integer 1000)] [level(integer 95)] [eform(string)]
```

## -bts- et -stbts-

varlist(numeric): la liste des variables indépendantes.

[if]: voir le manuel de Stata.

[in]: voir le manuel de Stata.

cmd(string): le nom de l'instruction que l'on veut utiliser, par exemple -reg-, -logit- ou -ologit-.

[cmdops(string)]: s'il y a lieu, les options de cette instruction, telles qu'elles sont présentées dans la section appropriée des manuels de Stata.

PWeight(varname numeric): le poids d'échantillonnage conventionnel.

rw(varlist numeric): la liste des variables qui contiennent les poids de rééchantillonnage.

[r(integer 1)]: s'il y a lieu, le nombre d'échantillons rééchantillonnés dont on a fait la moyenne pour générer **des poids de rééchantillonnage moyens**. Par exemple, 25 dans le cas de l'ESG 2001.

[dof(integer 1000)]: le nombre de degrés de liberté sur lequel seront basés les tests. En principe, ce nombre devrait être égal au nombre d'individus utilisés pour estimer le modèle divisé par un estimé de l'effet de plan moyen. En cas de doute, on peut utiliser le nombre par défaut que j'ai choisi. On peut également supprimer ce paramètre du PROGRAMME -stbts- (i.e. effacer [dof(integer 1000)] de l'instruction -syntax- et effacer les références à la variable locale `dof' dans la suite du PROGRAMME -bts-).

[level(integer 95)]: la largeur des intervalles de confiance, par défaut 95%.

[eform(string)]: pour obtenir les coefficients sous forme exponentielle (v.g.. de rapports de risque pour -stcox-). Par défaut, ils sont affichés sous forme additive.

## -bts- et -stbts-

### Exemples:

```
xi, prefix(): bts incm i.relig6, cmd(ologit)
  rw(wtbs*) r(25) dof(3103) eform(Rap cotes)
  pw(wght_per)
```

```
xi, prefix(): stbts QC37A-OC37A OA37A-00800
  i.NivelEduc, cmd(stcox) rw(wtbs*) r(25)
  dof(3103) eform(Rap risque) pw(wght_per)
```

# Exemples d'utilisation

- Voir les fichiers
  - prg1 (1b).do et
  - prg1 (1b).log

## Limites de -bts- et -stbts-

- Ne permet pas de calculer la variance par rapport à la moyenne des estimés obtenus par rééchantillonnage, mais seulement par rapport aux estimés obtenus à partir de l'échantillon original.

# Programmes

<http://www.uqs.inrs.ca/Cours/laplante/Longitudinal.htm>